GAN-based Pose-aware Regulation for Video-based Person Re-identification

GAN-based Pose-aware Regulation for Video-based Person Re-identification

Alessandro Borgia[1,2], Yang Hua[3], Elyor Kodirov[4], Neil M. Robertson[3]
[1]Heriot-Watt University, [2]University of Edinburgh, [3]Queen's University Belfast, [4]Anyvision
ab41@hw.ac.uk,{y.hua,n.robertson}@qub.ac.uk, elyor@anyvision.co

Abstract

Video-based person re-identification deals with the inherent difficulty of matching unregulated sequences with different length and with incomplete target pose/viewpoint structure. Common approaches operate either by reducing the problem to the still images case, facing a significant information loss, or by exploiting inter-sequence temporal dependencies as in Siamese Recurrent Neural Networks or in gait analysis. However, in all cases, the inter-sequences pose/viewpoint misalignment is not considered, and the existing spatial approaches are mostly limited to the still images context. To this end, we propose a novel approach that can exploit more effectively the rich video information, by accounting for the role that the changing pose/viewpoint factor plays in the sequences matching process. Specifically, our approach consists of two components. The first one attempts to complement the original pose-incomplete information carried by the sequences with synthetic GAN-generated images, and fuse their feature vectors into a more discriminative viewpoint-insensitive embedding, namely Weighted Fusion (WF). Another one performs an explicit pose-based alignment of sequence pairs to promote coherent feature matching, namely Weighted-Pose Regulation (WPR). Extensive experiments on two large video-based benchmark datasets show that our approach outperforms considerably existing methods.

1. Introduction

In surveillance, person re-identification (re-id) has emerged as a fundamental capability that no tracking system aiming to operate over a wide area network of disjoint cameras can, concretely, renounce to have. Recently, the person re-id task has broadened from the still images context [44, 42, 53, 2, 1] to video-based approaches, either supervised [52, 30, 64, 58, 47, 13] or semi-unsupervised [57, 26, 54]. For the video-based person re-id, given a query sequence of images (a.k.a., tracklet) of the target of interest, the challenge consists of identifying all the corresponding matching tracklets captured across the network. Dealing with full sequences of images as opposed to single images offers significant advantages: a) exploiting the temporal dependencies between *intra*-sequence frames [52, 30, 56, 71]; b) extracting more robust spatial appearance descriptors [64, 51, 58]; c) partially recovering from occlusions and reducing the influence of the background [46, 47, 13]. All these three aspects contribute to reducing the impact of the factors affecting performance (changing pose/viewpoint, lighting, occlusions, etc.), due to the availability of more diverse samples of the same identity. Depending on whether the frame-level feature aggregation scheme is based on temporal cues, estimated noise-state of frames, motion-state or pose/viewpoint (that is spatial cues), we can identify different *alignment* methods for matching corresponding video-fragments features (see Sec.2.2 for related references).

Temporal alignment methods usually suffer from noisy unregulated sequences with severe viewpoint and lighting variability [47]. Under such scenarios, it is difficult to estimate accurately gait phase and cycle or the optical flow energy profile as in [61]. Furthermore, the background clutter and occlusions from other people (in crowded context) interfere with the target feature map computation, causing over-fitting. Differently, other approaches, disregarding inter-frames temporal dependencies, focus on building more robust spatial features by performing an overall average pooling operation across the entire feature maps sequence [64, 51, 69, 58], thus reducing the multiple feature maps to a single instance. Despite performance benefit from it, we argue that this summarizing operation is not performed optimally due to the following reasons: a) it is frame-indiscriminate and ignores the pose/viewpoint information; b) real sequences are typically pose-incomplete, therefore the extracted features may not be as discriminative as under the availability of complete pose-information. A few techniques targeting occlusion impact reduction divide the incomplete unaligned sequences according to the estimated noise-state of the frames [13] or their motion-state (from the optic flow intensity profile) [46, 47], or weight

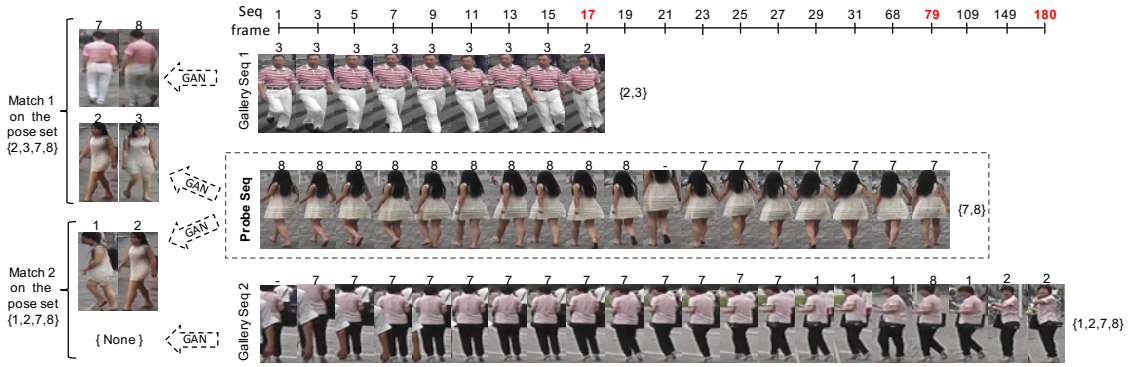

Figure 1. MARS [64] sequences with different length and unregulated pose/viewpoint at corresponding time frames. The probe sequence (in the middle) has to be matched against all the sequences of the gallery (top and bottom). For each (probe, test) sequence pair, the missing poses are produced and the corresponding images are generated by a GAN and added to the original sequences. For the sake of space constraints, the frame sequence axis is not linear. From the top to the bottom: IDs= $\{23, 1, 967\}$. Best viewed in color.

samples by a quality-aware network [25].

Our work originates from the observation that none of the mentioned approaches aiming to exploit the extra information of video-data account for the relevance of the viewpoint information in the sequence pairs matching process, despite the pose/viewpoint problem has been largely investigated in the image-based re-id [65, 67, 32, 27, 36]. We embrace the view that the need of accounting for pose/viewpoint is even greater in video-based re-id than in still images-based re-id. As proved in [58, 64], the ambiguity in distinguishing between video-based person representations can increase with respect to image-based representations, because the extra information associated with the motion is hardly discriminative. In order to incorporate the pose/viewpoint information in the sequences matching process for video-based person re-id, we propose a Pose-Driven Sequence Regulation (**PDSR**) approach articulated in two complimentary sub-schemes, the Weighted Fusion (WF) scheme and the Weighted Pose-Regulation (WPR) scheme. The first one enriches each sequence with more varied and complete viewpoint information by adding synthetic images, which are generated by a Generative Adversarial Network (GAN) with conveniently defined canonical poses (similarly to [32] for still images). This helps to synthesize people representations which generalize better to unknown identities, as shown in Figure 1. The second one (WPR) explicitly aligns sequences pairs by canonical poses to match spatial information coherently. Both schemes, differently from [23], operate at testing time, conceptually similar to the viewpoint repositioning strategy in [37] at recognition time. In summary, the main contributions of this paper are the following:

– To the best of our knowledge, this is the first work to apply a GAN-based generative model to video-based person re-identification for complementing and pose-aligning the original incomplete data.

– We identify the importance of pose-based coherent matching of sequences to exploit more effectively the available video-information and synthesize feature vectors with improved generalization capability.

– Our approach achieves a significant performance boost on two large video-based person re-id benchmarks with comparison to recently proposed techniques.

2. Related Work

2.1. Cross-View Invariant Techniques

The importance of accounting for the pose/viewpoint invariance problem, in person re-id, has been amply proved by many works. A popular approach is metric learning [61, 58, 72, 22, 51, 48] where a similarity metric is learned in the space of the video-level feature vectors expressing different views, aiming to increase the intra-class compactness and the inter-class distance of the identities. A different stream of research, complementary to metric learning, tackles the viewpoint problem by focusing on designing/learning more robust feature representations, for example, exploiting the temporal aggregation of multiple frame-level features maps [30] or performing spatial fusion/concatenation of global/local features [56, 4].

The introduction of GANs has represented a step forward in the way that the viewpoint problem is tackled because it allows creating new synthetic data under the desired viewpoints without requiring any extra labelling [32, 27, 38, 60, 62]. This is why it lends itself particularly well to be combined with the existing video-based feature extraction techniques for producing even more discriminative embeddings. In this paper, we follow this combined approach for complementing the incomplete pose information

of video-sequences and refer to [38] for the GAN architecture and to [51] for the feature extraction CNN.

2.2. Video-Sequences Alignment Techniques

Video-based person re-id demands to deal with the *intra*-sequence appearance variability caused by many changing factors across cameras (occlusions, pose and viewpoint, etc.), in addition to handling the visual ambiguity at the *inter*-matching items level as it happens in the still images case. To tackle this, a stream of literature has emerged proposing temporal [39, 9, 30, 56], spatio-temporal [24, 29], pose/viewpoint-based [41, 63, 45, 32, 50], motion-state-based [46, 47] and noise-state-based [13] sequence alignment techniques. Gao *et al.* [9] proposes a temporally aligned pooling representation for video-based person re-id which relies on dividing a sequence into several segments according to the sinusoid of a person walking cycle and then performing segment-based pooling to extract a representation. The periodicity of pedestrians gate is exploited also in [24] to generate a spatio-temporal body-action model made up of a series of action primitives of certain body parts treated independently from each other. The strict alignment assumptions made by gait recognition techniques are relaxed in [47, 61] where unregulated video-sequences are automatically broken down based on motion energy profiling (e.g., optical flow). With the spreading of the deep learning paradigm, temporal-based approaches have emerged using RNNs (combined with Convolutional Neural Networks) in Siamese configuration [30, 56, 52, 71]. These methods embed both the mid-term temporal information in the video-sequence representation and the long-term appearance information summarized by the temporal pooling layer. Another form of sequence alignment is the spatial regulation, handling the body part misalignment problem across frames, usually addressed by defining part-based body models, either pre-defined (stripes or grids) [5, 53, 42] or learned from data by spatio-temporal attention mechanisms [41, 63, 55, 21] that take into account the non-rigid shape of the human body.

2.3. GANs for Person Image Generation

Generative Adversarial Networks (GANs) [10] represent the most popular approach to deep learning-based generative image modelling, compared to either the unsupervised techniques based on variational autoencoders [18] or autoregressive models [31]. A distinction in two classes of GANs can be made depending on the nature of the input distribution to the generator: noise-driven GANs [68, 28] where a mapping from a Gaussian noise distribution to the images distribution is learned and attribute conditional GANs [16]. As to the former class, [28] represents for the person re-id task a notable work where a disentangled person representation with respect to pose, background and foreground is learned. Zheng *et al.* [68] applies DCGAN [33] to the person re-id task to improve the discriminative learning by assigning a uniform label distribution of the generated images over all the existing classes. Therefore, it assumes that the generated data belong to none of the training classes, differently from [35] where the synthetic new samples are treated as a single extra class. In our work, instead, following [32], we assign identity membership labels to the generated images in viewpoint-based canonical forms, leveraging the good quality of the synthetic data, instead of using them as a regularization strategy for outliers. The noise-based GANs formulation is less effective in person re-id due to the challenging nature of the task that requires to condition the generated images to some attributes like pose [27, 32, 62, 38], clothing [19], camera style [70], dataset style [49, 7] for domain adaptation. We adopt the conditional architecture proposed in [38] because of its capability to handle the global morphological transformation happening between the input-target poses pair.

3. Proposed Method

Our approach consists of considering the pose/viewpoint role, in terms of both completeness and alignment. In order to match together a pair of sequences, we convert each of them into a viewpoint-normalized form made up of a set of pre-defined canonical poses manually selected from the train set. As to the pose-information completeness aspect, the basic idea of our technique is that before matching two sequences they should be enriched with complete appearance information of the observed target. In other words, they should contain instances of the entire set of possible discrete viewpoints in which the pedestrian may be shot. Integrating this information in all sequences allows extracting better informative deep representations of the identities which in turn translates into a more successful features matching. The viewpoint instances of a target identity, which are not present in the original data, need to be generated. This is done by conditional GANs that can generate synthetic images of pedestrians in any desired pose [38, 32, 68, 60, 43, 28, 14, 62, 27] and dataset style [49, 17]. We accomplish the integration of the missing viewpoint information according to the Weighted Fusion approach that finds a workable balance between the contribution of the original sequences to the final embedding and the contribution of the synthetic GAN-generated images.

As to the sequences pose-based alignment aspect, having pose-misaligned frames causes incoherent sequences matching, which results in sub-optimal performance. To address this, we propose the Weighted-Pose Regulation method based on explicitly aligning the sequence pairs based on pose/viewpoint. The two techniques, WF and WPR, leverage the integration with three state-of-the-art deep frameworks: the light-weight residual learning-based

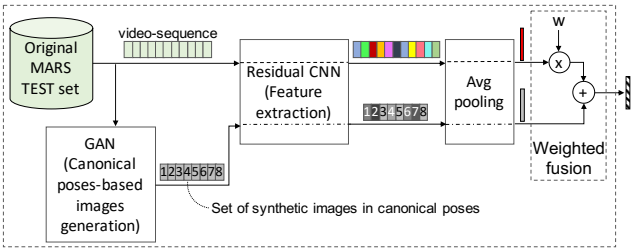

Figure 2. Illustration of the Weighted Fusion strategy. Best viewed in color.

CNN model in [51] (our baseline) for frame-level feature extraction; the key-points detector [3] for detecting the human joints, also used in [27, 38]; the deformable GAN model in [38] for its ability to learn the global human body transformations across cameras. We design a balanced combination of these separate modules into a unified video-based person re-id framework and study its overall and ablative performance.

3.1. Weighted Fusion (WF)

The WF scheme (Figure 2) is based on aggregating (by average pooling) the frame-level features of a real sequence separately from the frame-level features of the corresponding generated sequence of canonical poses, in order to exploit the complementarity of their contributions. Finally, the two embeddings are fused together into a more discriminative representation. Using canonical poses, similarly to [32], allows reducing the continuous variability of the real observable poses to a discretized finite pose-space for affordable pose-to-pose matching. Differently from all other papers, we operate directly on the testing set images without fine-tuning the network weights on the generated sequences in canonical form and investigate the two contributions balance.

Given a video-based training dataset $D_{Tr} = \{\tilde{\mathbf{x}}_i\}$ and a testing set $D_{Ts} = \{\mathbf{x}_i^{(s)}\}$, with $\mathbf{x}_i^{(s)} \in \mathbb{R}^d$ denoting the i-th image of the s-th video sequence, we firstly select from D_{Tr} a set of M images of pedestrians $C = \{\tilde{\mathbf{x}}_{c_j}\}_{j=1}^{M}$, $c_j \in \{1, ..., |D_{Tr}|\}$ representative of M significant canonical viewpoints in which the pose space can be quantized (Figure 3, bottom row). The different viewpoints are intended as angular rotations around the imaginary longitudinal axis drawable across the human body. Secondly, for each $\tilde{\mathbf{x}}_{c_j} \in C$, the key-points detector proposed in [3] is used to extract a 2D coordinates vector of k joints and mapping it to a pose-image $\mathcal{P}(\tilde{\mathbf{x}}_{c_j})$ that summarizes the key-points information and that we refer to as canonical image (Figure 3, middle row). Thirdly, the heat maps $\mathcal{H}(\mathcal{P}(\tilde{\mathbf{x}}_{c_j}))$ of the canonical poses, representing the pose-conditioning input to the GAN, are calculated. Fourthly, for each testing sequence, one image $\mathbf{x}_i^{(s)}$ representative of the pedestrian identity is randomly drawn and used to generate a set of M

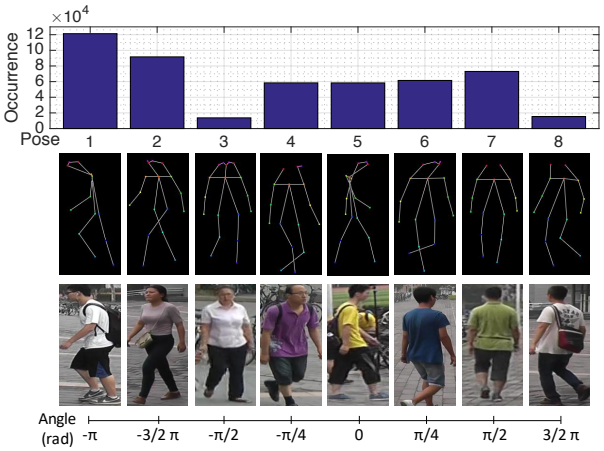

Figure 3. Illustration of the canonical poses on the MARS dataset. Top: Canonical poses distribution across the MARS dataset. Middle: The 8 canonical pose-images $\mathcal{P}(\tilde{\mathbf{x}}_{c_j})$ that summarize the key-points information of the correspondent 8 manually selected images $\tilde{\mathbf{x}}_{c_j}$ from MARS (bottom). Best viewed in color.

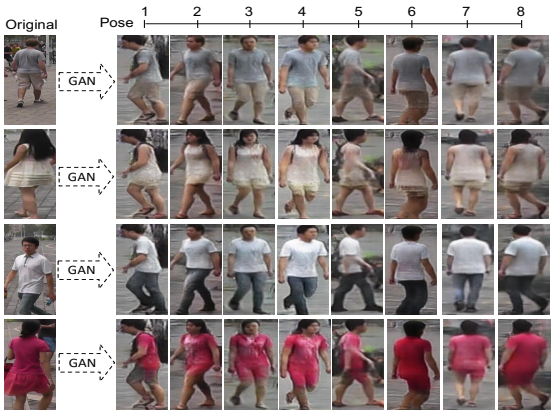

Figure 4. Translation of four MARS identities into sequences of GAN-generated images in canonical form (From the top: IDs = {262, 1, 507, 1205}). Best viewed in color.

synthetic images $\{\hat{\mathbf{x}}_{i,j}^{(s)}\}_{j=1}^{M}$ with the same identity as $\mathbf{x}_i^{(s)}$ but represented under the M canonical poses (Figure 4). For the generation of these images three inputs to the GAN are required: $\mathbf{x}_i^{(s)}$ with its associated pose-image $\mathcal{H}(\mathcal{P}(\mathbf{x}_i^{(s)}))$ (as detailed in [38]) and the heat maps of the canonical poses. This generation process creates a new synthetic test-set $\hat{D}_{T_s}$ corresponding, sequence-by-sequence, to the original one D_{T_s}, which is exploitable to extract video-level features $\mathbf{h}^{(s)}$ more robust from the viewpoint invariance perspective. The features extracted from the s-th test sequence can be mathematically expressed as in Equation 1:

$$\mathbf{h}^{(s)} = w \cdot [\frac{1}{L} \bigoplus_{i=1}^{L} f(\mathbf{x}_i^{(s)})] \oplus [\frac{1}{M} \bigoplus_{j=1}^{M} f(\hat{\mathbf{x}}_{i,j}^{(s)})] \quad (1)$$

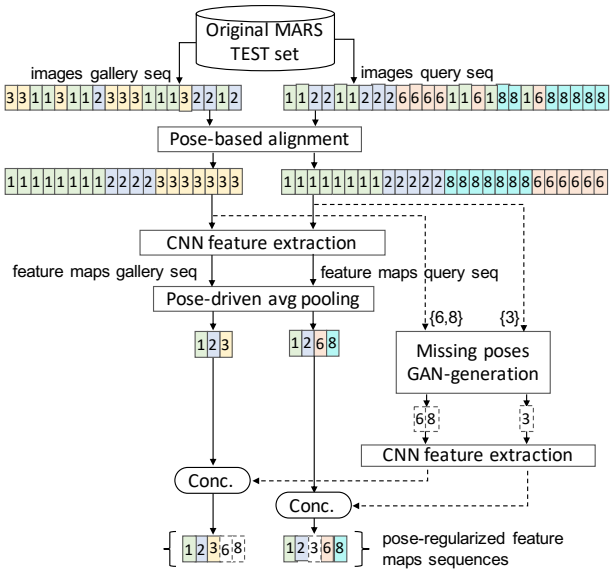

Figure 5. Pose-based sequence pair normalization block.

where $\bigoplus$ indicates the Kronecker element-wise sum operation of multiple matrices (feature maps); $\oplus$ denotes the element-wise sum of two matrices; w is the weighting parameter of the WF method; L is the length of the original sequence that $\mathbf{x}_i$ belongs to; $f(\cdot)$ represents the feature extraction function of the CNN learned on the training set.

The synthetic image $\hat{\mathbf{x}}_i$ can be represented as in Equation 2:

$$\hat{\mathbf{x}}_i = G(\mathbf{z}, \mathbf{x}_i, \mathcal{H}_i, \tilde{\mathcal{H}}_j) \tag{2}$$

where G is the generator of the GAN; $\mathbf{z}$ is a noise vector that is implicitly incorporated by the network dropout; $\mathcal{H}_i \equiv \mathcal{H}(\mathcal{P}(\mathbf{x}_i))$ and $\tilde{\mathcal{H}}_j \equiv \mathcal{H}(\mathcal{P}(\tilde{\mathbf{x}}_j))$ are the heat maps of a generic image and of one canonical pose image belonging to C, respectively. The generator optimization is driven by two losses, the standard conditional adversarial loss $\mathcal{L}_{GAN}(G, D)$ and the nearest-neighbour loss $\mathcal{L}_{NN}(G)$ defined in [38], as in Equation 3

$$G^* = \arg\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{NN}(G) \tag{3}$$

where D is the GAN discriminator and λ the coefficient of the linear combination. It is noteworthy to underline that the weighting parameter w (explored throughout its space of variation) plays the critical role of controlling the level of noise introduced into the deep representation by the generated images, in order to make effective fusion.

3.2. Weighted Pose-Regulation (WPR)

The sequence pair alignment method based on the viewpoint factor represents an independent source of performance improvement. We propose a scheme articulated into two blocks: one block performs viewpoint-based sequence pair normalization (Figure 5), and another deals with the actual feature vectors matching (Figure 6). As to the sequence normalization block, given a pair of tracklets X_{s_1} and X_{s_2} of arbitrary length to be matched, each of them is reassembled into up to M view-point specific sub-sequences $X_s = [X_s^{(p_1)}, ..., X_s^{(p_j)}]$, $1 \leqslant j \leqslant M$, grouping images that share the same canonical pose $p_j \equiv \mathcal{P}(\tilde{\mathbf{x}}_{c_j})$. With regards to the canonical poses introduced in Section 3.1, one image $\mathbf{x}_i^{(s)}$ is considered to have the j-th canonical pose if the Euclidean distance of its extracted key-points vector from it is the smallest, that is: $|\mathcal{P}(\mathbf{x}_i^{(s)}) - \mathcal{P}(\hat{\mathbf{x}}_{c_j}^{(s)})| \leqslant |\mathcal{P}(\mathbf{x}_i^{(s)}) - \mathcal{P}(\hat{\mathbf{x}}_{c_l}^{(s)})|$, $\forall j \neq l$, $j, l \in \{1, ..., M\}$, with $|\cdot|$ denoting the Euclidean distance. By performing frame-level feature extraction $f(\cdot)$ and average pooling $avg(\cdot)$ separately on each single pose-specific sub-sequence $X_s^{(p_j)}$, a video-based representation $\mathbf{f}^{(s)} \equiv [avg(f(X_s^{(p_1)})), ..., avg(f(X_s^{(p_j)}))] \equiv [\mathbf{f}^{(s,p_1)}, ..., \mathbf{f}^{(s,p_j)}]$ is produced as the concatenation of view-specific contributions $\mathbf{f}^{(s,p_j)}$. At this stage, the video-level multi-pose aggregated representations $\mathbf{f}^{(s_1)}$ and $\mathbf{f}^{(s_2)}$ of the two sequences X_{s_1} and X_{s_2} are not yet pose-aligned, because each one includes a different sub-set of canonical poses $R^{(s)} \equiv \{p_j\}_{j \in \{1,...,M\}}$. In order to align them, they both must include, respectively in $\mathbf{f}^{(s_1)}$ and $\mathbf{f}^{(s_2)}$, one video-feature map for each of the canonical poses in $R^{(s_1)} \cup R^{(s_2)}$. The missing ones are GAN-generated and concatenated. Denoting by $Q^{(s)} \equiv \{p_m\}_{m \in \{1,...,M\}}$ the set of the missing poses in $\mathbf{f}^{(s)}$, with $Q^{(s)} \cap R^{(s)} = \emptyset$ and $Q^{(s)} \cup R^{(s)} = C$, we can express the final pose-normalized sequence representation as in Equation 4

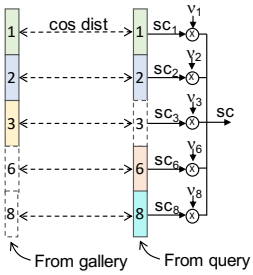

Figure 6. Matching strategy. One pair of pose-aligned sequences of feature maps are matched pose by pose by the cosine distance.

$$\overline{\mathbf{h}}^{(s)} = cat_{sort}([\mathbf{f}^{(s)}, ..., \{\mathbf{f}^{(s,p_m)}\}_{m \in Q^{(s)}}]) \tag{4}$$

where $cat_{sort}(\cdot)$ is a function that performs vertical concatenation of the $\mathbf{f}^{(s,p_j)}$ by sorting all terms according to the increasing pose index $j = 1, ..., M$. With regards to the matching/ranking block, we design an ad-hoc strategy for

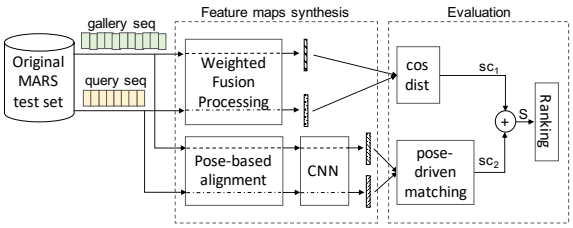

Figure 7. Overall system architecture. The pose-driven matching block is shown in Figure 6, the Pose-based alignment block in Figure 5 and the weighted fusion processing in Figure 2.

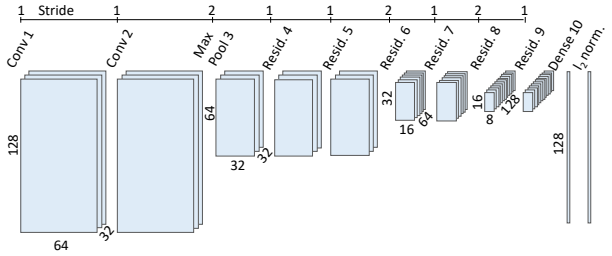

Figure 8. Residual learning-based CNN architecture [51] used for feature extraction. The patch size for the convolutional, pooling and residual layers is 3×3.

matching pairs of pose-normalized feature map sequences $\overline{\mathbf{h}}^{(s_1)}$, $\overline{\mathbf{h}}^{(s_2)}$, based on performing separate sub-matching of pose-specific feature maps $\{\mathbf{f}^{(s,p_m)}\}$ according to how illustrated in Figure 6, where the weight parameter ν_j represents the averaged value of the frequencies with which the canonical pose p_j occurs in the two matching sequences.

3.3. Combining WF and WPR

Due to the conceptual complementarity of the two methods, we can combine them together, shown in Figure 7, to benefit from both additive contributions at the same time. The fusion happens at the score level: the two score vector produced by WF and WPR (using the cosine distance) are summed up element-wise and ranked by decreasing overall score. With this approach, the final ranking list accounts for both pose information completeness and pose alignment, promoting progress in the ranking for those feature matchings that would get penalized by the single WF score calculation because of their frame-based pose misalignment.

4. Experiment

4.1. Database

MARS (Motion Analysis and Re-identification Set) [64] is one of the largest video re-id datasets currently available and represents an extension of Market-1501 [66]. It contains $1,261$ IDs (631 belonging to the training set and 630 to the testing set) captured by at least two of the 6 cameras deployed. The $20,715$ video-sequences of MARS are automatically generated by the Deformable Part Model pedestrian detector [8] and the GMMCP tracker [6]. MARS reproduces the challenges of a real-world scenario, due to the presence of $3,248$ distractors and of many partial and total occlusions that make its tracklets quite noisy.

DukeMTMC-VideoReID [54] is a large-scale video-based re-identification dataset, created from the multi-target multi-camera tracking dataset DukeMTMC [34], that counts $1,404$ valid IDs captured by up to 8 disjoint static cameras plus 408 IDs which are distractors appearing in only one camera. The pedestrians are randomly split into a training set and a testing set, each one with 702 IDs. The training set is structured in $2,196$ tracklets made up of 369,656 frames, while the testing set includes $2,636$ tracklets formed by $445,764$ frames.

4.2. Evaluation Metrics and Protocols

The evaluation protocol of MARS and DukeMTMC-VideoReID datasets are borrowed from the Market1501 dataset [66] with the difference that, for the video-based datasets, probes and queries are tracklets instead of still images. For each identity, one query tracklet representing that identity is randomly selected and tested against the gallery sequences which are finally ranked based on their cosine similarity to the query. The overall performance is calculated by averaging across all identities. The gallery sets of the probes include all the sequences belonging to camera views different to the probe, including distractors. The metrics that we use for evaluating our person re-id algorithms against both datasets are the Cumulated Matching Characteristics (CMC) curve and the mean-Average Precision (mAP) in order to account jointly for precision and recall. Furthermore, in order to measure the relative improvement of re-id between camera pairs, we build the mAP camera-pairs confusion matrix, shown in Figure 10, for the WF+WPR method on MARS against the DCML cosine-softmax baseline [51]. For a fixed pair of cameras (X, Y), where X is the probe camera and Y represents the test camera, the corresponding value in the confusion matrix is calculated by limiting the positive samples (tracklets) to only those ones viewed under camera Y and averaging for all the probes viewed under the same camera X.

4.3. Implementation Details

Feature extraction CNN. The CNN architecture that we use for feature extraction is the one proposed in [51], shown in Figure 8, based on the residual learning scheme [11] and principles [59]. It is a quite shallow network made up of only 15 layers in order to be suitable for the task of people tracking which requires fast features extraction capabilities. Dropout [40] and batch normalization [15] are ap-

Table 1. State-of-the-art (%) on the MARS. $^{(a)}$ indicates a training/architectural advantage due to the use of Imagenet pre-training. $^{(b)}$ denotes semi-unsupervised methods.

Method	Rank 1	mAP
CAR + Video [61]	55.5	-
ETAP Net [54] $^{(a,b)}$	62.67	42.45
IDE (C) + XQDA [64] $^{(a)}$	65.30	47.60
IDE (R) + ML [69]	70.51	55.12
CaffeNet [71]	70.60	50.70
MSCAN [20]	71.77	56.06
IDE (R) + ML [12]	72.42	57.42
DCML (baseline) [51]	72.93	56.88
P-QAN [25]	73.73	51.70
Ours	**75.76**	**60.57**

plied between layers for regularization purposes. The input is represented by RGB images rescaled to 128×64. The cosine-softmax classifier replaces the standard softmax classifier in order to get more compact classes in the feature space representation [51]. The CNN is trained directly on the target re-id dataset with a batch size of 128 images. For training the network for the identity classification task, 631 classes for MARS and 702 for DukeMTMC-VideoReID are set and the output of the l_2 normalization layer (i.e., a 128 elements vector) is used as feature embedding for the following matching step.

The training is regulated by a learning rate $\eta = 10^{-3}$ and the network was regularized by a weight decay of 10^{-1} and a dropout with probability 40%. The model checkpoint at 46461 iterations is selected as the best one.

GAN model. For generating synthetic images, we borrow the deformable GAN presented in [38] because it bypasses the commonly required two stages training that is replaced by a single stage end-to-end training. The generator and the discriminator are trained on the Market-1501 dataset for $90,000$ iterations, with batch size equal to 4 and warping skip layers of type "mask" that allows applying the warping only to the foreground target removing the background, like in [27]. The optimization is performed by the Adam optimizer with learning rate $\alpha = 2 * 10^{-4}$, exponential decay rate for the first moment $\beta_1 = 0.5$ and for the second moment $\beta_2 = 0.999$. We train the GAN on Market-1501 instead of directly on MARS because the latter represents a more noisy video-based extension of the Market-1501.

Pose estimation model. For pose estimation we use the real-time multi-person model (Human Pose Estimator, HPE) proposed in [3] and also used in [38, 27] that is based on a non-parametric representation of 18 landmarks corresponding to the human body joints locations.

Table 2. State-of-the-art (%) on the DukeMTMC-VideoReID dataset. $^{(a)}$ denotes training/architectural advantage due to the use of Imagenet pre-training and a deeper network. $^{(b)}$ denotes semi-unsupervised methods.

Method	Rank 1	mAP
DGM+IDE [57] $^{(a)}$	42.36	33.62
Stepwise [26] $^{(a)}$	56.26	46.76
ETAP Net [54] $^{(a,b)}$	72.79	63.23
Ours	**82.22**	**78.76**

4.4. Comparison with State-of-the-art Methods

Our method, relying on a light-weight CNN, outperforms most of the state-of-the-art techniques in video-based person re-id on MARS (Table 1) and on DukeMTMC-VideoReID (Table 2). With regards to the MARS dataset, we achieve a rank 1 accuracy of 75.76% and a mAP of 60.57% with an improvement respectively of $+2\%$ and $+8.9\%$ with respect to the second best method [25]. For the sake of completeness, we should clarify that we exclude from the comparison a few methods that are not directly comparable with ours because they use more powerful learning strategies with more data, or deeper and more complex architectures (Siamese): one examples is [12], excluded from Table 1 because it employs a triplet network, although, despite that, our results are still higher than those of its LuNet network ($rank_1 = 75.56\%$ and $mAP = 60.48\%$). Another examples is [21] that uses ResNet50 pre-trained on 6 re-id datasets and fine-tuned on three additional ones with overall four stages training.

Also on the DukeMTMC-VideoReID dataset, both WF and WPR succeed in boosting the performance of our baseline, reaching the state-of-the-art on this very new dataset (Table 2). We achieve an improvement of $+9.4\%$ and $+15.5\%$ respectively for rank 1 accuracy and mAP, with regards to the second best performing method, ETAP Net [54]. For a fair comparison, it should be pointed out that [54] is a semi-unsupervised technique while ours is a supervised one. By the way, ETAP Net provides also a supervised upper bound (83.62% and 78.34% for rank 1 accuracy and mAP respectively) which is achieved with a much deeper network, ResNet50, pre-trained on the Imagenet dataset. Despite our method relies on a network with a depth of around $1/3$ of ResNet50 and is trained directly on the target train set, compared to the ETAP Net supervised upper bound, we reduce the architecture structure disadvantage and achieve competing results (-1.4% $+0.4\%$ for rank 1 accuracy and mAP respectively). Our results show that the features extracted from GAN-generated images with a model learned only on the real images can be used for identity discrimination at testing time other than for network regularization at training time as in [68], maintaining the identity label of their original conditioning images. This

Table 3. Ablation study on MARS w.r.t WF and WPR (%).

Method	Rank 1	mAP
DCML$^{(*)}$ (baseline)	72.57	56.57
WF (ours)	74.19 (+1.6)	59.48 (+2.9)
WF + WPR (ours)	**75.76 (+3.2)**	**60.57 (+4)**

Table 4. Ablation study on DukeMTMC-VideoReID (%).

Method	Rank 1	mAP
DCML$^{(*)}$ (baseline)	79.20	75.25
WF (ours)	80.84 (+1.6)	77.56 (+2.3)
WF + WPR (ours)	**82.22 (+3)**	**78.76 (+3.5)**

makes our approach easy to be applied to different models for getting a boost in performance with no need of retraining. It is worth noting that adding the M synthetic canonical poses to each identity sub-sequence in the *train set* for data augmentation, does not improve the performance. Furthermore, it negatively affects the re-id, even for small values of the augmentation factor. We reckon that this is due to the high level of noise introduced into the learning process by the less discriminative synthetic images which make the learned video-level features more ambiguous. Exploiting GAN-generated images for data-augmentation in person re-id has been done in [70] for camera style adaptation. A substantial difference between the camera style transfer case and ours, though, is that while in [70] the synthetic images are conditioned on the camera style and thus preserve the real content of the original conditioning images, in our work the conditioning is done on pose which is a more difficult characteristic to reproduce while preserving the identity, because it involves an image content change.

4.5. Ablation Analysis

On DukeMTMC-VideoReID, both WF and WPR improve the baseline (+79.20% and 75.25% respectively for rank 1 accuracy and mAP) that simply performs the average pooling of the frame-level features. Table 4 shows that 53% of the rank 1 accuracy improvement (66% for the mAP) is due to the WF method, while the remaining +47% (34% for mAP) to the WPR viewpoint-based alignment technique, which highlights the effectiveness of our two schemes and demonstrates that they are complementary to some extent.

With regards to the baseline [51], on MARS, our PDSR technique allows an overall improvement of +3.2% and +4% respectively for the rank-1 accuracy and the mAP Table 3. This improvement is due in part to the weighted fusion strategy WF (+1.6% and +2.9% over the baseline, respectively, for rank 1 accuracy and mAP) and for the remaining part to the WPR technique (+1.6% and +1.1% for rank-1 accuracy and mAP). With regards to WF, we report in Figure 9 the curve describing how the rank-1 accuracy varies over the weighting parameter space of the linear

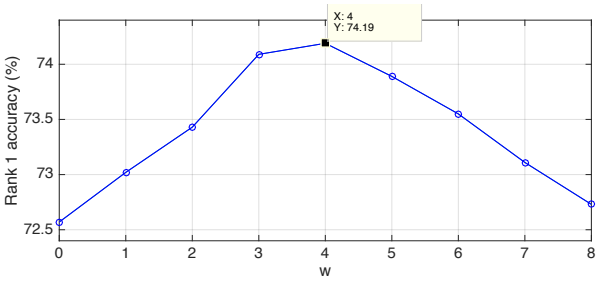

Figure 9. Rank 1 accuracy on MARS. Baseline @$w = 0$.

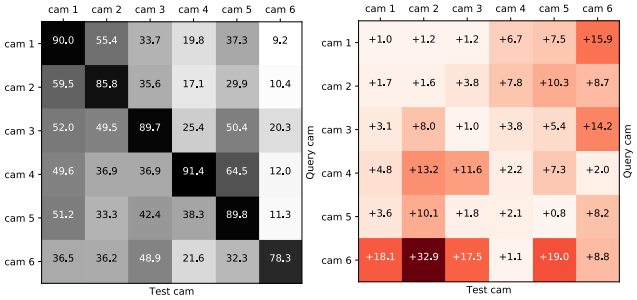

Figure 10. Left: mAP confusion matrix (%) for DCML [51]. Right: mAP relative confusion matrix for WF+WPR w.r.t. DCML.

combination. The top value, corresponding to $w = 4$, defines the WF performance. Evidence of the benefit of our method to mitigate the effects of the inter-camera viewpoint problem when matching pedestrians tracklets comes from the mAP confusion matrix in Figure 10 for it shows that the highest relative improvements to the re-id performance happen for the cross-camera matchings compared to the intra-camera re-id cases (confusion matrix main diagonal).

5. Conclusions

In this paper, with regards to the video-based person re-id task, we addressed the problem of how to account effectively for the pose/viewpoint information in the pose-incomplete and noisy video-sequences matching process at the testing time. We formulated two separate approaches, WF and WPR, that work also jointly and rely on the definition of canonical poses, weight-controlled fusion, generated canonical poses sequences and viewpoint-based sequences alignment. Combined, our techniques achieve state-of-the-art performance on MARS and DukeMTMC-VideoReID.

Acknowledgement

This work was supported in part by the UDRC Consortium, University Defence Research Collaboration in Signal Processing and in part by Roke Manor Research.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] A. Borgia, Y. Hua, E. Kodirov, and N. M. Robertson. Cross-view discriminative feature learning for person re-identification. *IEEE Transactions on Image Processing*, 27:5338–5349, 2018.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[4] L. Chen, H. Yang, J. Zhu, Q. Zhou, S. Wu, and Z. Gao. Deep spatial-temporal fusion network for video-based person re-identification. In *CVPRW*, 2017.

[5] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[6] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, 2015.

[7] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[9] C. Gao, J. Wang, L. Liu, J.-G. Yu, and N. Sang. Temporally aligned pooling representation for video-based person re-identification. In *ICIP*, 2016.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[12] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[13] W. Huang, C. Liang, Y. Yu, Z. Wang, W. Ruan, and R. Hu. Video-based person re-identification via self paced weighting. In *AAAI*, 2018.

[14] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang. Multi-pseudo regularized label for generated samples in person re-identification. *arXiv*, 2018.

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[17] N. Jetchev and U. Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *ICCV*, 2017.

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[19] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. In *ICCV*, 2017.

[20] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[21] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.

[22] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[23] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *CVPR*, 2018.

[24] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *ICCV*, 2015.

[25] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017.

[26] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, 2017.

[27] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NIPS*, 2017.

[28] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *CVPR*, 2018.

[29] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[30] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016.

[31] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016.

[32] X. Qian, Y. Fu, W. Wang, T. Xiang, Y. Wu, Y.-G. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *ECCV*, 2018.

[33] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[34] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016.

[36] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. *CoRR*, abs/1711.10378, 2017.

[37] S. Savarese and L. Fei-Fei. View synthesis for recognizing unseen poses of object classes. In *ECCV*, 2008.

[38] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable gans for pose-based human image generation. *arXiv*, 2017.

[39] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *ECCV*, 2012.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural

networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[41] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee. Part-aligned bilinear representations for person re-identification. *arXiv*, 2018.

[42] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[43] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017.

[44] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016.

[45] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *arXiv*, 2018.

[46] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[47] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514, 2016.

[48] X. Wang, Y. Hua, E. Kodirov, G. Hu, and N. M. Robertson. Deep metric learning by online soft mining and class-aware attention. In *AAAI*, 2019.

[49] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. *arXiv*, 2017.

[50] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACM Multimedia*, 2017.

[51] N. Wojke and A. Bewley. Deep cosine metric learning for person re-identification. In *WACV*, 2018.

[52] L. Wu, C. Shen, and A. v. d. Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *CoRR*, abs/1606.01609, 2016.

[53] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016.

[54] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018.

[55] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[56] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, 2017.

[57] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, 2017.

[58] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *CVPR*, 2016.

[59] S. Zagoruyko and N. Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[60] C. Zhang, L. Wu, and Y. Wang. Crossing generative adversarial networks for cross-view person re-identification. *arXiv*, 2018.

[61] W. Zhang, S. Hu, and K. Liu. Learning compact appearance representation for video-based person re-identification. *CoRR*, abs/1702.06294, 2017.

[62] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng. Multi-view image generation from a single-view. In *ACM Multimedia*, 2018.

[63] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[64] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.

[65] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *CoRR*, abs/1701.07732, 2017.

[66] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[67] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *CoRR*, abs/1707.00408, 2017.

[68] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[69] Z. Zhong, L. Zheng, D. Cao, and S. L. R.-r. P. Re. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[70] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.

[71] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 2017.

[72] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *IJCAI*, 2016.